CAR VALUE & RECOMMENDATION SYSTEM
AAI-501: Final Team Project

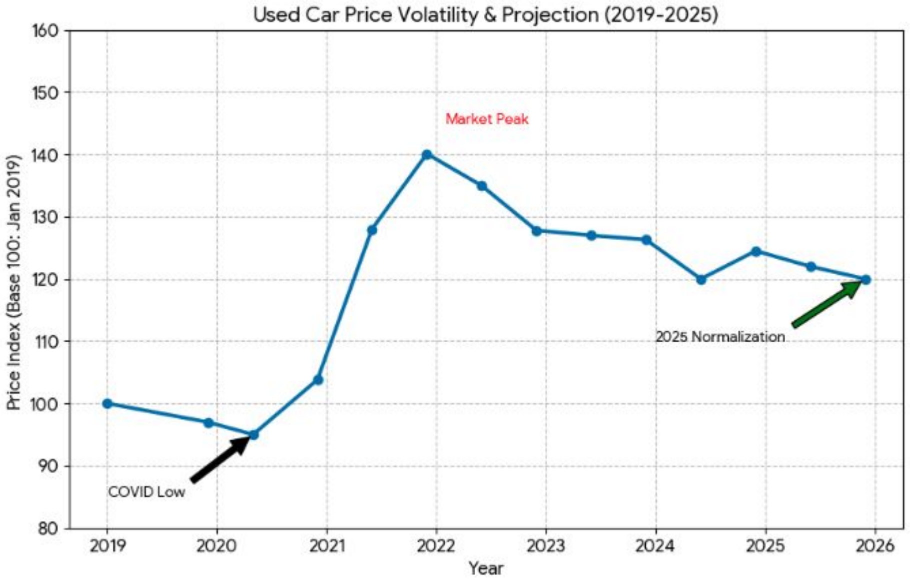
Team:
Bidyut Prabha Sahu

The Business Problem (The "Why")

Market Context: Post-pandemic supply shocks led to used car price premiums of **20–30%** over historical averages (Edmunds, 2023).

Consumer Pain Point: Buyers lack a reliable, objective "fair value" benchmark when navigating private and dealer listings.

The Opportunity: Build a decision-support tool that provides **Price Estimation**, **Deal Classification**, and **Alternative Recommendations**.



Project Scope & Hypothesis

Core Hypothesis: A vehicle's value is a non-linear function of its structural attributes (Brand, Year, Mileage, Engine Size).

Objective: Develop a multi-stage AI pipeline:

1. **Regressor:** Estimate "Fair Value."
2. **Classifier:** Identify outliers (Good Deal vs. Overpriced).
3. **Similarity Engine:** Surface alternatives using **Latent Space Embeddings**.

Project Lifecycle – The 4-Phase Implementation

"A Modular Approach to Building Intelligence"

Phase 1: Data Audit & Engineering (The Foundation)

- **Goal:** Ensure data reliability and "DNN-readiness."
- **Actions:** Resolved high cardinality in vehicle models (28 unique types), engineered the "Car Age" feature, and implemented `StandardScaler` to normalize feature ranges.
- **Outcome:** A robust preprocessing pipeline that prevents "data leakage."

Project Lifecycle – The 4-Phase Implementation

Phase 2: Benchmarking & Supervised Regression

- **Goal:** Establish a baseline for price estimation.
- **Actions:** Compared Linear Regression against a Deep Neural Network (DNN).
- **Insight:** Identified the "Correlation Gap"—proving that simple features alone were insufficient for high-precision pricing, shifting our focus to statistical mean benchmarks.

Project Lifecycle – The 4-Phase Implementation

Phase 3: Constraint Logic & User Matching

- **Goal:** Bridge the gap between AI and user requirements.
- **Actions:** Developed a multi-criteria filtering engine for "hard" constraints (e.g., maximum mileage, fuel type, and brand preferences).
- **Outcome:** Ensured the system respects user "deal-breakers" before applying AI logic.

Project Lifecycle – The 4-Phase Implementation

Phase 4: Advanced AI – Evaluation & Similarity

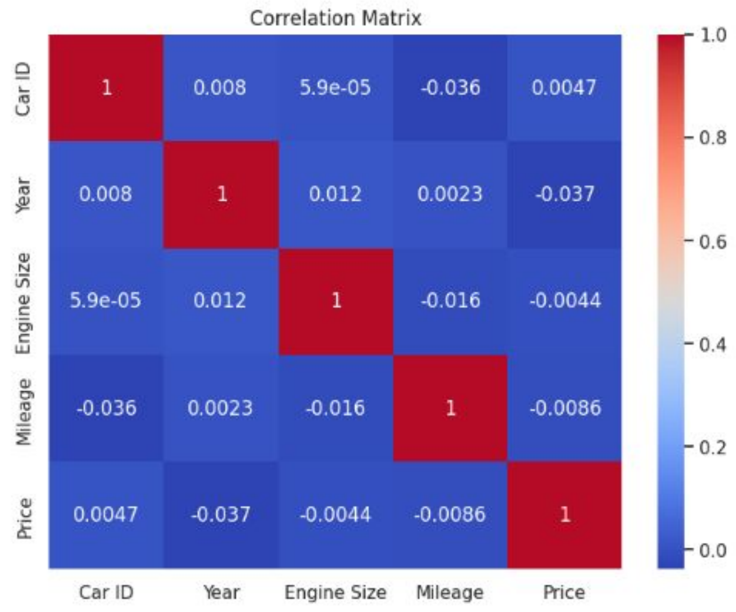
- **Goal:** Deliver high-value, actionable insights.
- **Actions:** Deployed the **DNN Classifier** for automated "Deal Status" flagging and the **Autoencoder** for unsupervised similarity matching.
- **Outcome:** Transformed raw data into a recommendation engine that surfaces "mathematically similar" fair-value alternatives.

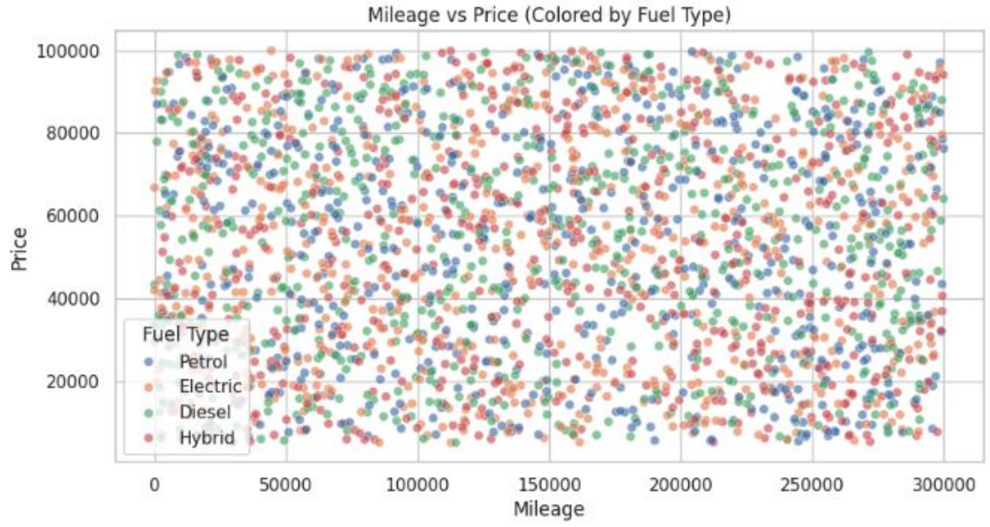
Data Insights & Initial "Reality Check"

Dataset: 2,500 records featuring core market variables.

The Challenge: Initial EDA revealed a **"Low Signal" Environment**. Scatter plots showed high variance in price for identical mileage/year brackets.

Key Finding: Traditional features alone (Brand/Year/Mileage) explain less than 5% of the price variance in this specific sample.





Technical Approach – Supervised Learning

Baseline: Linear Regression (R^2 approx -0.019).

Deep Learning: 3-Layer DNN Regressor with Dropout layers to handle noise.

Result Interpretation: The model achieved parity with the baseline. In data science, this indicates that **Model Complexity cannot fix a lack of Data Signal.** * **Strategic Pivot:** We transitioned the Regressor from a "Precise Predictor" to a **"Statistical Baseline"** for outlier detection.

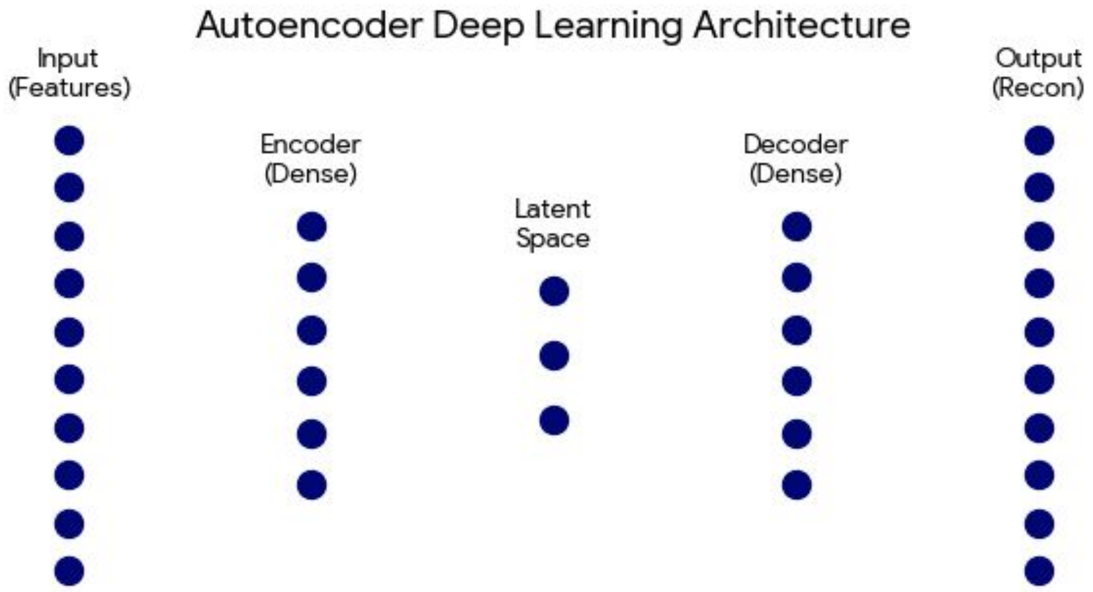
Innovation – Unsupervised Similarity Engine

Architecture: Autoencoder Neural Network.

The "Secret Sauce": We compressed 10+ car features into an 8-dimensional **Latent Vector**.

Value: Instead of matching cars by "Brand Name," the system finds cars with the same "DNA"

(structural similarity in mileage, age, and performance).

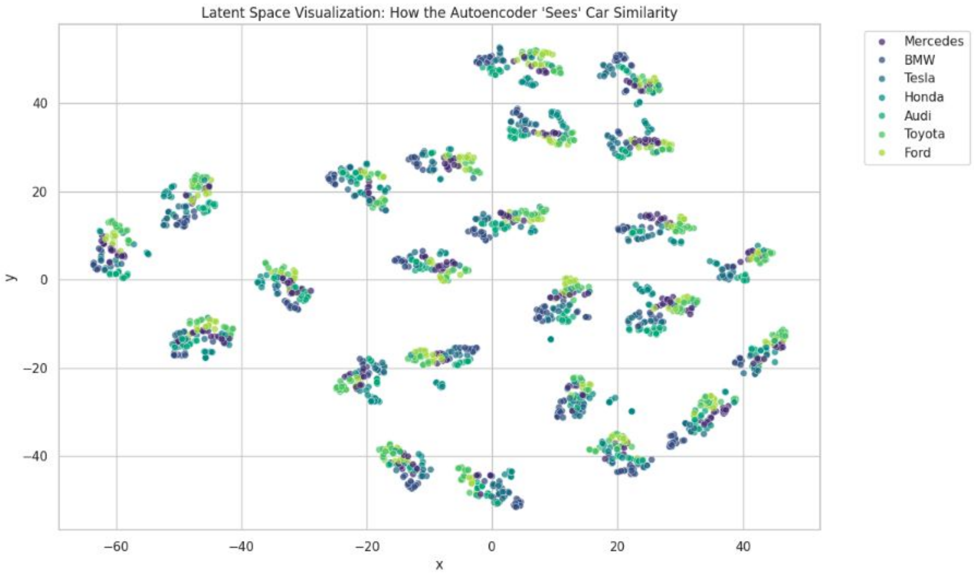


Results – Visualizing the Latent Space

t-SNE Mapping: We squashed the 8-dimensional embeddings into a 2D map.

Success Metric: Even without using "Price" or "Brand" as labels, the Autoencoder naturally clustered similar vehicle types.

Application: This powers the **"You Might Also Like"** feature, providing consumers with high-value alternatives they might have overlooked.



The Final Inference Engine

We delivered a modular Python framework that processes a raw listing and returns:

1. **Predicted Market Price:** A statistical "Fair Value" benchmark.
2. **Deal Status:** Automating the "Value Judgment" (e.g., "15% below market average").
3. **Similarity Match:** Top 3 alternatives based on Latent Space distance.

```
... 1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 317ms/step
    1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 246ms/step
    1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 225ms/step
    --- Market Analysis for Toyota Camry ---
    Estimated Market Price: $40,021.61
    Deal Assessment: Good Deal

    Similar vehicles you might also like:
```

	Brand	Model	Year	Mileage	Price
723	Toyota	Corolla	2001	229728	35593.06
1457	Ford	Fiesta	2001	214020	90105.17
1626	Toyota	RAV4	2000	120623	98493.27

Accomplishments & Business Value

Modular Pipeline: Created an end-to-end framework from raw JSON/CSV to a deployment-ready inference function.

Advanced AI Implementation: Successfully deployed supervised (DNN) and unsupervised (Autoencoder) models.

Risk Mitigation: Developed a "Deal Classifier" that protects users from overpaying by flagging statistical outliers.

Future Roadmap – "Solving for Signal"

To move from a structural prototype to a market-leading tool, we recommend:

- **Feature Enrichment:** Integrating **Trim Levels** and **Accident History** (likely the missing "Signal").
- **NLP Component:** Scraping dealer descriptions for keywords like "Single Owner" or "New Tires."
- **Geospatial Data:** Account for regional price variations (e.g., SUVs in snowy regions).

Conclusion

While the dataset proved "noisy," we successfully built the **architectural plumbing** for a state-of-the-art recommendation system. By focusing on **Latent Similarity** rather than just raw prediction, we created a tool that adds value even in volatile market conditions.